FLOWVQA: MAPPING MULTIMODAL LOGIC IN VISUAL QUESTION ANSWERING WITH FLOWCHARTS

Shubhankar Singh[1†], Purvi Chaurasia[2†], Yerram Varun[3*], Pranshu Pandya[4*], Vatsal Gupta[4*], Vivek Gupta[5‡], Dan Roth[5]

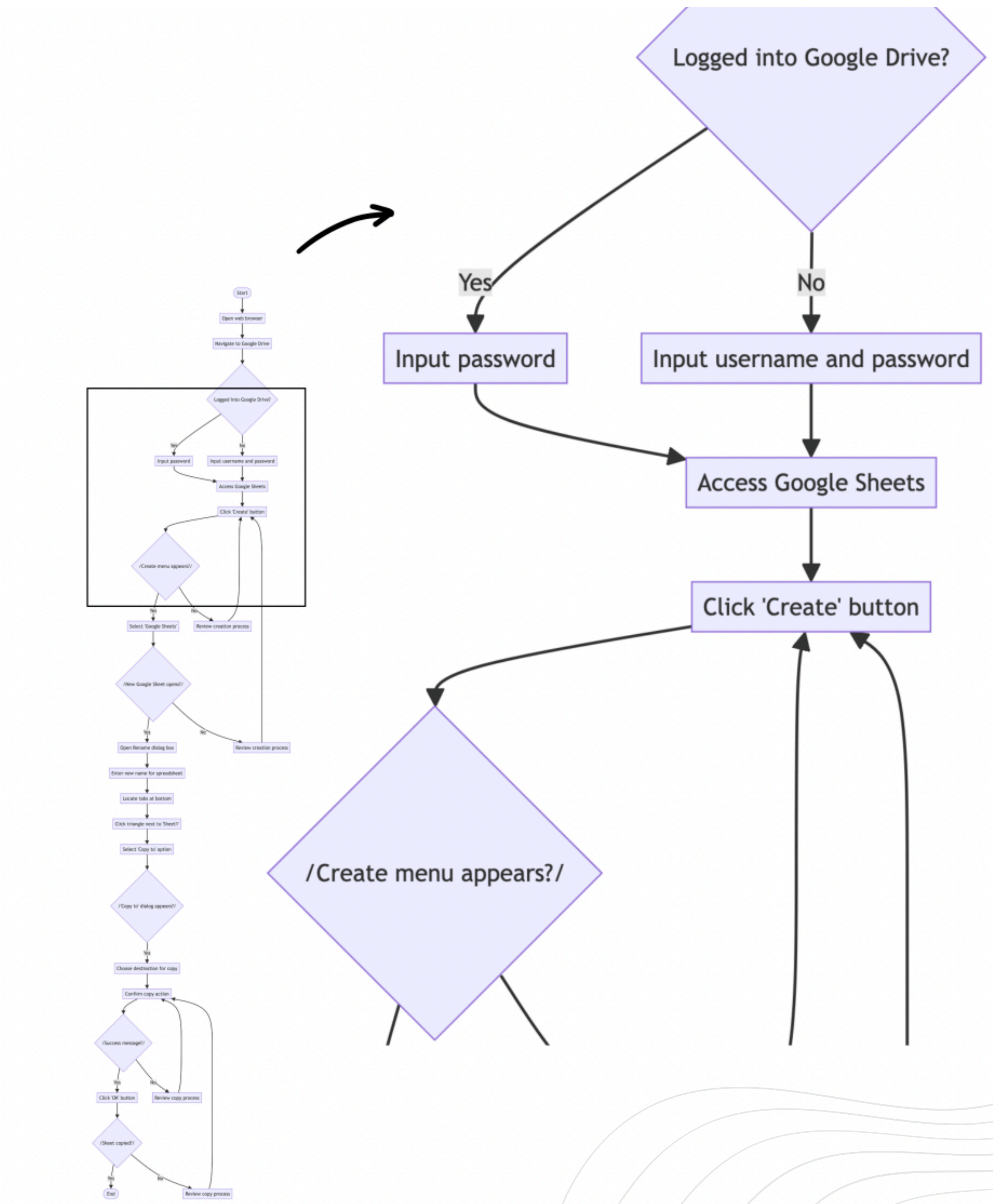
[1] Mercer Mettl, [2] IGDTUW New Delhi, [3] Google Research, [4] Indian Institute of Technology Guwahati, [5] University of Pennsylvania

ACL 2024

Bangkok, Thailand

PROBLEM DEFINITION

- Existing Visual Question Answering (VQA) benchmarks lack emphasis on visual grounding and complex spatial reasoning.
- They often do not assess models' abilities in understanding intricate visual structures like flowcharts.
- ***FlowVQA:*** 2,272 flowchart images sourced from various instructional and technical content. 22,413 diverse pairs target reasoning skills such as information localization, scenario deduction, and logical progression.



Q. Derek wants to ensure that the sheet was successfully copied before reporting back to Melissa. What should Derek see or do next to ensure the task was completed correctly?

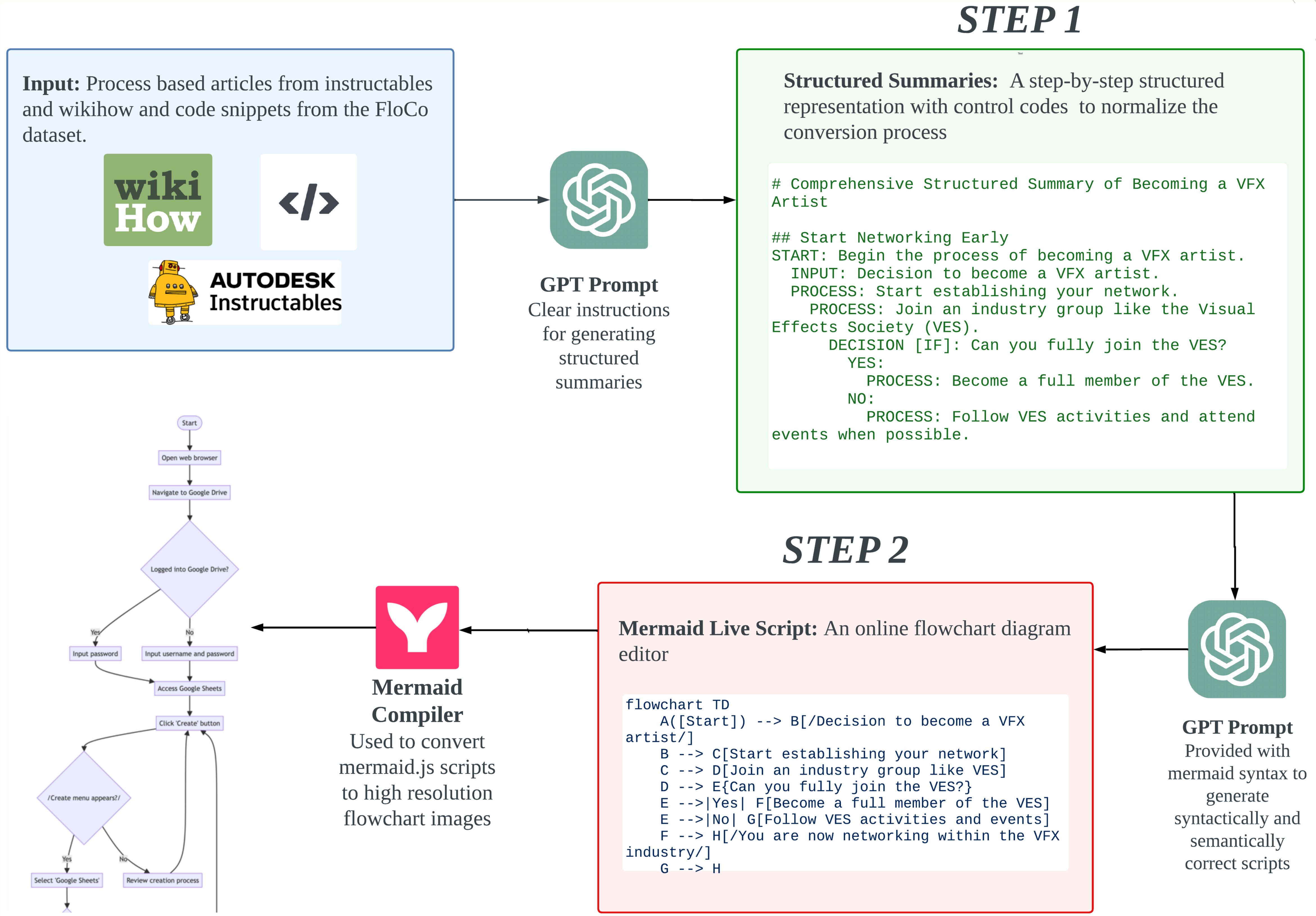
A. He should look for a success message and dismiss the dialogue by clicking ’OK’.

ACL 2024

Bangkok, Thailand

DUAL STEP FLOWCHART GENERATION PROCESS

- GPT-4 is used to generate structured representations from source texts, such as converting instructional articles into step-by-step guides.
- **First Step:** Source texts are converted into a structured format with control tags (e.g., "START," "PROCESS," "DECISION") to outline the process flow.
- **Second Step:** The structured text is transformed into Mermaid.js flowchart scripts using predefined templates, with control tags helping to map steps to specific node types.
- The scripts are compiled into high-resolution PNG images of flowcharts..



STEP 1

Input: Process based articles from instructables and wikihow and code snippets from the FloCo dataset.

GPT Prompt Clear instructions for generating structured summaries

Structured Summaries: A step-by-step structured representation with control codes to normalize the conversion process

```
# Comprehensive Structured Summary of Becoming a VFX Artist

## Start Networking Early
START: Begin the process of becoming a VFX artist.
  INPUT: Decision to become a VFX artist.
  PROCESS: Start establishing your network.
    PROCESS: Join an industry group like the Visual Effects Society (VES).
      DECISION [IF]: Can you fully join the VES?
        YES:
          PROCESS: Become a full member of the VES.
        NO:
          PROCESS: Follow VES activities and attend events when possible.
```

STEP 2

GPT Prompt Provided with mermaid syntax to generate syntactically and semantically correct scripts

Mermaid Live Script: An online flowchart diagram editor

```
flowchart TD
    A([Start]) --> B[/Decision to become a VFX artist/]
    B --> C[Start establishing your network]
    C --> D[Join an industry group like VES]
    D --> E{Can you fully join the VES?}
    E -->|Yes| F[Become a full member of the VES]
    E -->|No| G[Follow VES activities and events]
    F --> H[/You are now networking within the VFX industry/]
    G --> H
```

Mermaid Compiler Used to convert mermaid.js scripts to high resolution flowchart images

ACL 2024
Bangkok, Thailand

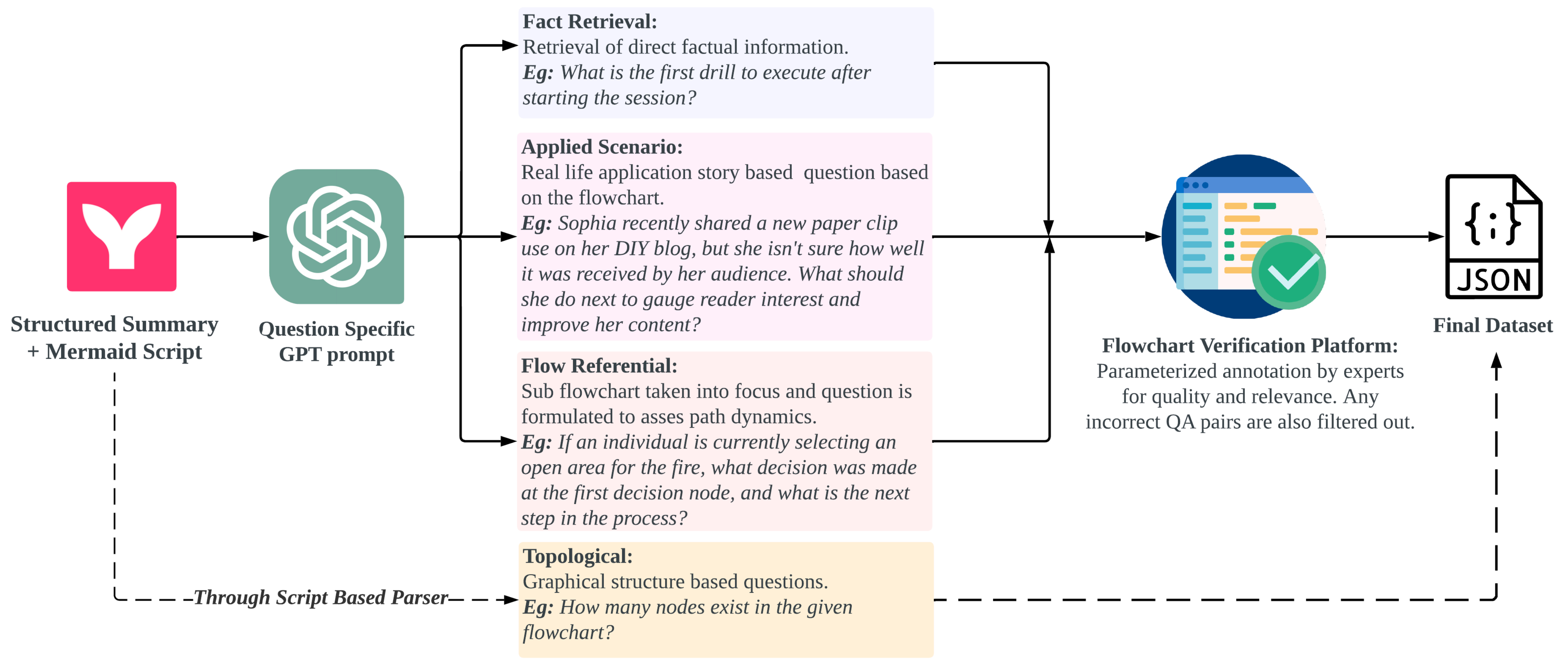
Q/A GENERATION AND VERIFICATION PROCESS

- Four types of questions are created: Fact Retrieval, Applied Scenario, Flow Referential, and Topological. These categories assess various aspects of flowchart comprehension and reasoning skills.

- GPT-4 is used to generate questions and answers based on tagged textual representations, Mermaid.js scripts, and few-shot examples. Each question type has specific prompts to ensure quality and relevance.

- Topological questions are created using a graph syntax parser and script.

- Each question has three paraphrased "gold standard" answers to accommodate variations in model responses. The Q/A pairs undergo a rigorous human verification process to ensure accuracy and quality.



Structured Summary + Mermaid Script

Question Specific GPT prompt

Fact Retrieval:
Retrieval of direct factual information.
Eg: *What is the first drill to execute after starting the session?*

Applied Scenario:
Real life application story based question based on the flowchart.
Eg: *Sophia recently shared a new paper clip use on her DIY blog, but she isn't sure how well it was received by her audience. What should she do next to gauge reader interest and improve her content?*

Flow Referential:
Sub flowchart taken into focus and question is formulated to asses path dynamics.
Eg: *If an individual is currently selecting an open area for the fire, what decision was made at the first decision node, and what is the next step in the process?*

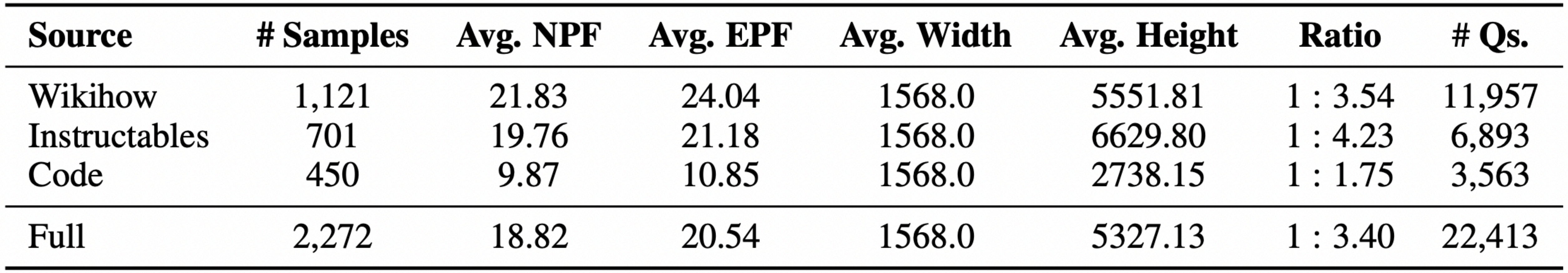
Topological:
Graphical structure based questions.
Eg: *How many nodes exist in the given flowchart?*

Through Script Based Parser

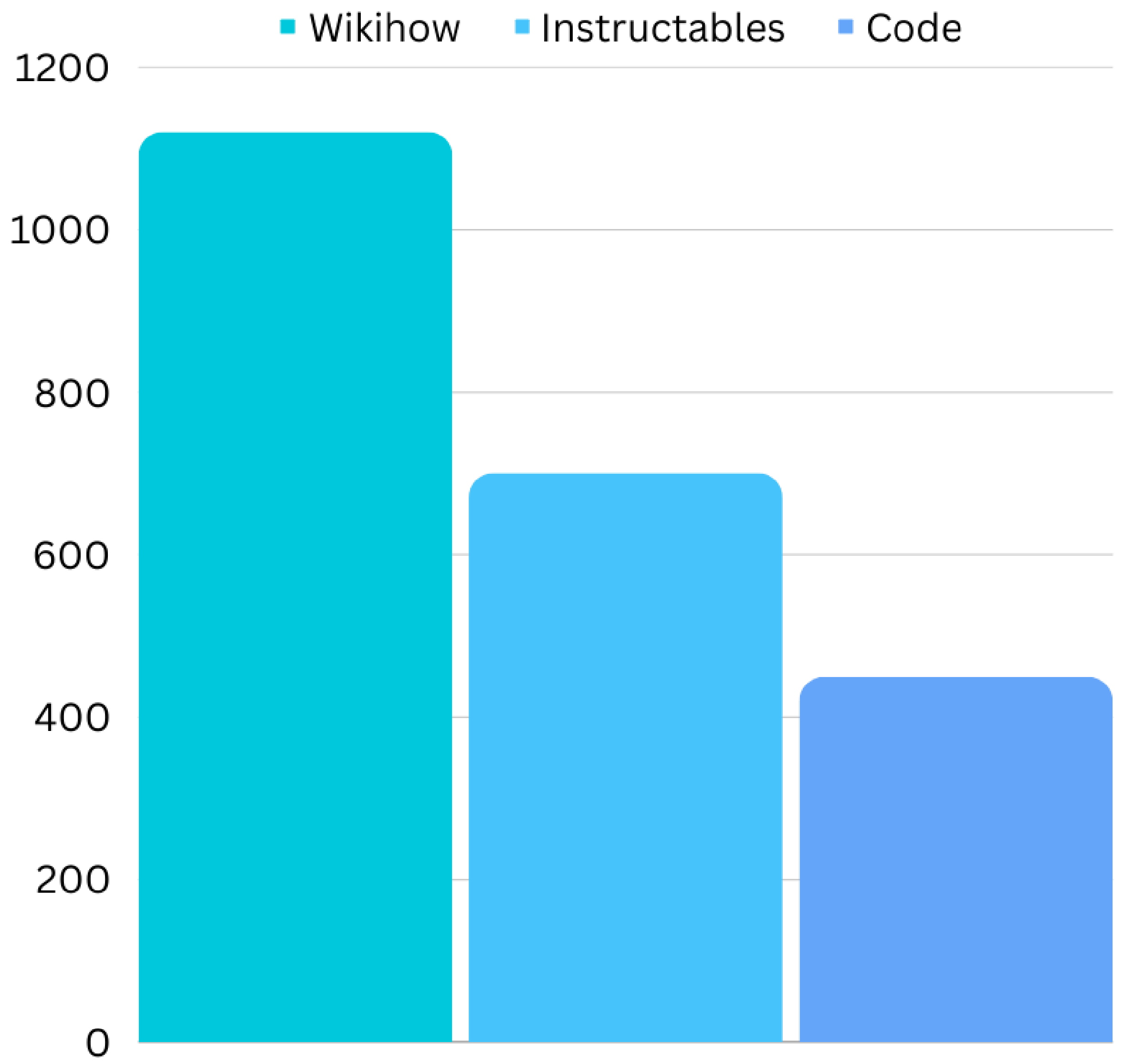
Flowchart Verification Platform:
Parameterized annotation by experts for quality and relevance. Any incorrect QA pairs are also filtered out.

Final Dataset

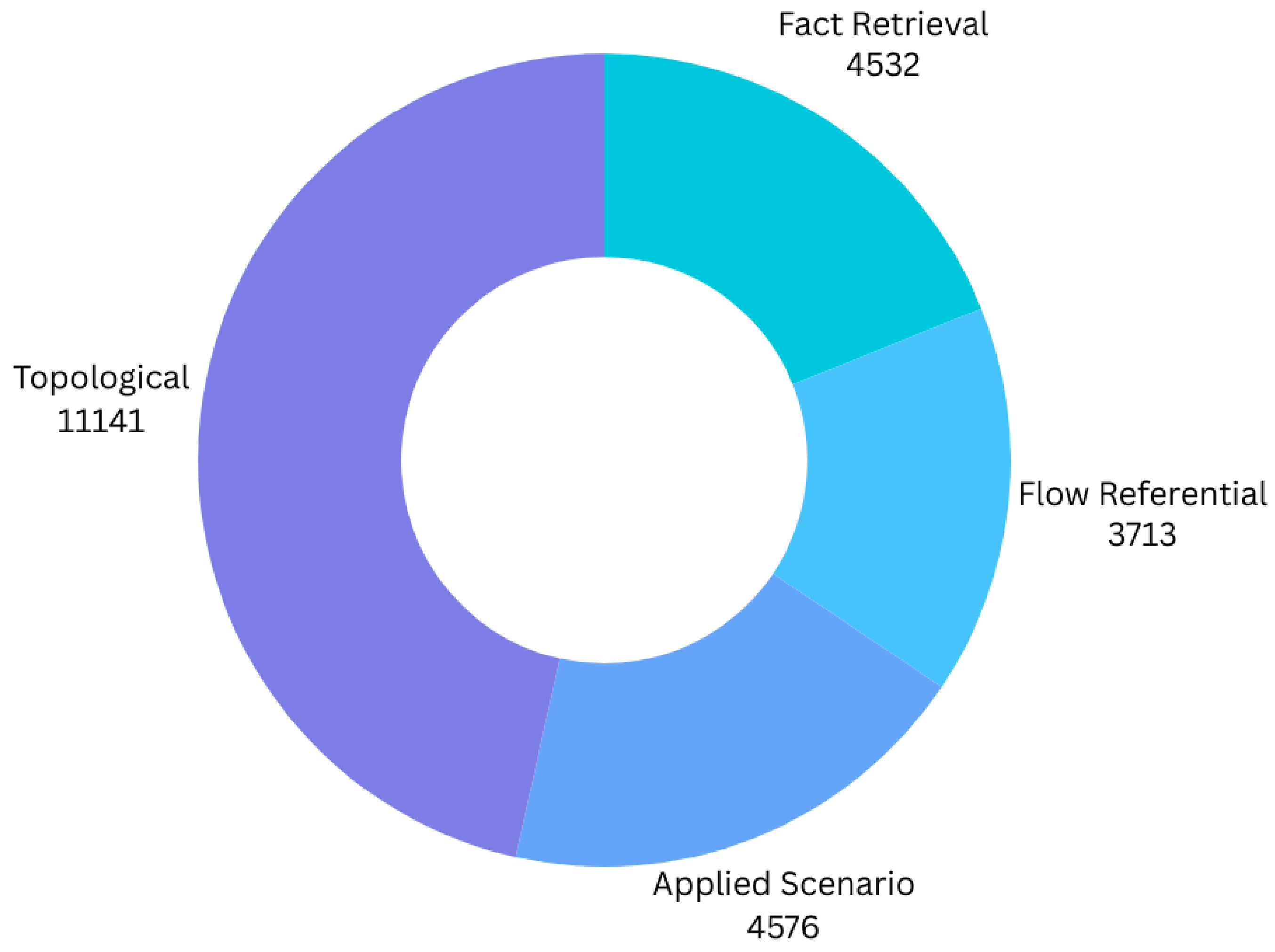
DATASET STATASTICS



Source	# Samples	Avg. NPF	Avg. EPF	Avg. Width	Avg. Height	Ratio	# Qs.
Wikihow	1,121	21.83	24.04	1568.0	5551.81	1 : 3.54	11,957
Instructables	701	19.76	21.18	1568.0	6629.80	1 : 4.23	6,893
Code	450	9.87	10.85	1568.0	2738.15	1 : 1.75	3,563
Full	2,272	18.82	20.54	1568.0	5327.13	1 : 3.40	22,413



Dataset distribution across source



Dataset distribution across Question Types

BASELINE EVALUATION SETTING

RQ1: Does the newly introduced visual multimodal dataset, FlowVQA, present a significant challenge to current multimodal language models (VLMs), and can it offer valuable insights for their future development?

RQ2: How do factors such as (a) the source of flowcharts, (b) the type of questions posed, and (c) the inherent complexity of the flowcharts affect the efficacy of VLMs?

RQ1: Can the performance of VLMs on visual question answering tasks related to flowcharts be enhanced through specific directives tailored to flowcharts? Additionally, does fine-tuning these models with the training split of the FlowVQA dataset improve their proficiency?

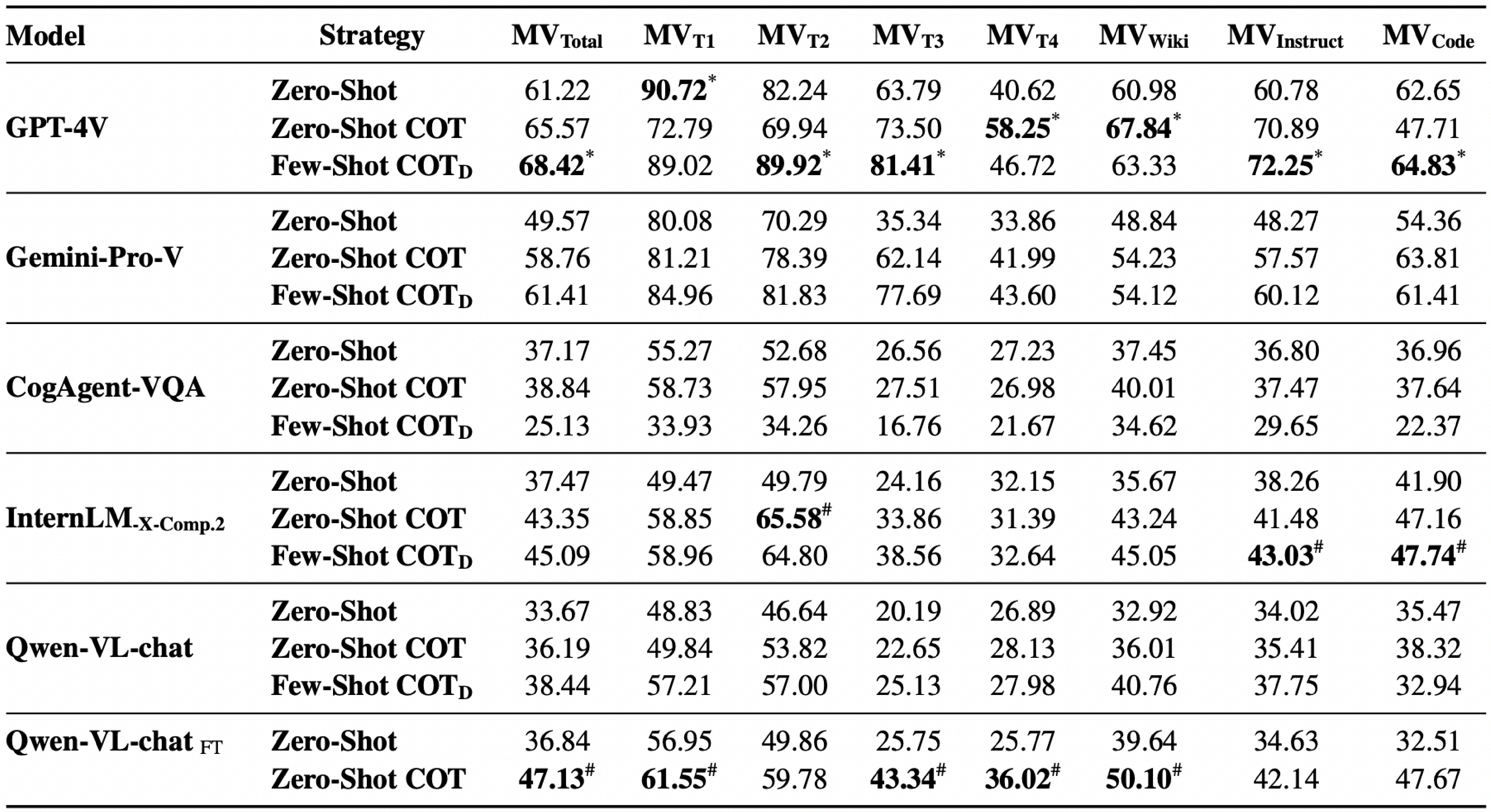
RQ4: Is there an observable directional bias in existing VLMs when handling flowchart-based visual question answering tasks?

- Setting for **Zero Shot and Zero-Shot Chain of Thought (CoT)** is same as other work. **Text-Only Few-Shot CoT with Reasoning Directives:** A custom prompt outlines reasoning steps specific to flowchart-related questions. This approach uses a few examples (few-shot) to guide the model through directional stimulus tags, step-by-step rationales, and answers.

- **Fine-Tuning:** The VLM is fine-tuned on the FlowVQA training set and then prompted to answer questions, enhancing the model's performance on the specific dataset.

- Responses generated by the VLMs are evaluated by three other models (GPT-3.5, Llama-2 70B, Mixtral 8*7B) to determine correctness through a detailed rationale and majority vote, focusing on the accuracy and coherence of the responses.

BASELINE RESULTS



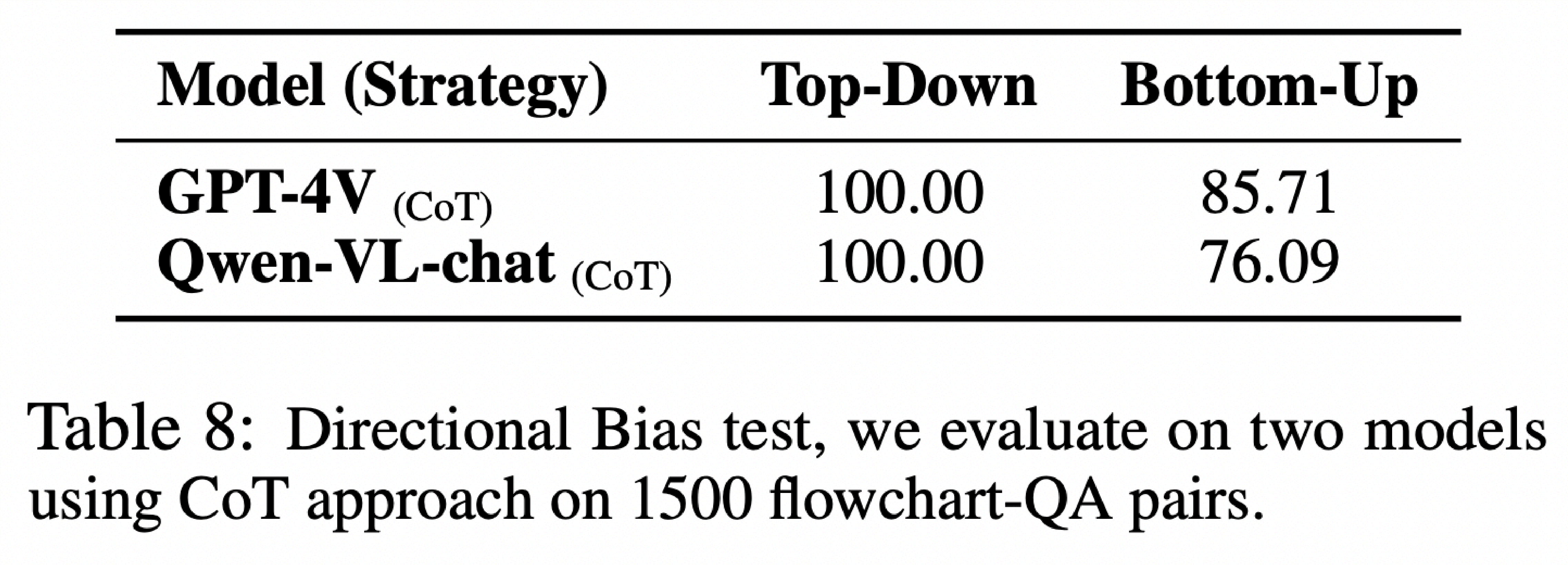
Model	Strategy	MV$_{\text{Total}}$	MV$_{\text{T1}}$	MV$_{\text{T2}}$	MV$_{\text{T3}}$	MV$_{\text{T4}}$	MV$_{\text{Wiki}}$	MV$_{\text{Instruct}}$	MV$_{\text{Code}}$
GPT-4V	**Zero-Shot**	61.22	**90.72***	82.24	63.79	40.62	60.98	60.78	62.65
	Zero-Shot COT	65.57	72.79	69.94	73.50	**58.25***	**67.84***	70.89	47.71
	Few-Shot COT$_{\text{D}}$	**68.42***	89.02	**89.92***	**81.41***	46.72	63.33	**72.25***	**64.83***
Gemini-Pro-V	**Zero-Shot**	49.57	80.08	70.29	35.34	33.86	48.84	48.27	54.36
	Zero-Shot COT	58.76	81.21	78.39	62.14	41.99	54.23	57.57	63.81
	Few-Shot COT$_{\text{D}}$	61.41	84.96	81.83	77.69	43.60	54.12	60.12	61.41
CogAgent-VQA	**Zero-Shot**	37.17	55.27	52.68	26.56	27.23	37.45	36.80	36.96
	Zero-Shot COT	38.84	58.73	57.95	27.51	26.98	40.01	37.47	37.64
	Few-Shot COT$_{\text{D}}$	25.13	33.93	34.26	16.76	21.67	34.62	29.65	22.37
InternLM$_{\text{X-Comp.2}}$	**Zero-Shot**	37.47	49.47	49.79	24.16	32.15	35.67	38.26	41.90
	Zero-Shot COT	43.35	58.85	**65.58**$^{\#}$	33.86	31.39	43.24	41.48	47.16
	Few-Shot COT$_{\text{D}}$	45.09	58.96	64.80	38.56	32.64	45.05	**43.03**$^{\#}$	**47.74**$^{\#}$
Qwen-VL-chat	**Zero-Shot**	33.67	48.83	46.64	20.19	26.89	32.92	34.02	35.47
	Zero-Shot COT	36.19	49.84	53.82	22.65	28.13	36.01	35.41	38.32
	Few-Shot COT$_{\text{D}}$	38.44	57.21	57.00	25.13	27.98	40.76	37.75	32.94
Qwen-VL-chat $_{\text{FT}}$	**Zero-Shot**	36.84	56.95	49.86	25.75	25.77	39.64	34.63	32.51
	Zero-Shot COT	**47.13**$^{\#}$	**61.55**$^{\#}$	59.78	**43.34**$^{\#}$	**36.02**$^{\#}$	**50.10**$^{\#}$	42.14	47.67

Table 6: Majority Vote Accuracy on All Models and Strategies broken down Question Type Wise (*T1, T2, T3, T4*) as in Sec 2.3 and Source-Wise (Instruct, Wiki, Code) as in Table 2. The highest value for each column is highlighted and marked with * in Closed Source Models and with # in Open Source Models.

DIRECTIONAL BIAS TEST

- Creating an inverted "Bottom Top" set of flowcharts, where start nodes are at the bottom and end nodes at the top, to test the VLMs' adaptability to non-standard flow directions.

- The top-performing models from earlier evaluations are tested on 1,500 inverted flowchart-question pairs to detect any directional bias, by comparing their performance on standard versus inverted flowcharts.



Model (Strategy)	**Top-Down**	**Bottom-Up**
GPT-4V (CoT)	100.00	85.71
Qwen-VL-chat (CoT)	100.00	76.09

Table 8: Directional Bias test, we evaluate on two models using CoT approach on 1500 flowchart-QA pairs.

DISCUSSION

- The dataset is challenging for all evaluated models, with the best-performing model achieving only 68.42% accuracy. This indicates a significant scope for improvement in handling complex visual information.

- Few-shot Chain of Thought (CoT) with reasoning directives significantly improves performance, particularly in proprietary models like GPT-4, which saw up to a 12% improvement compared to other strategies.

- Proprietary models generally outperform open-source models, with GPT-4 notably surpassing others by up to 30%. This highlights the potential for proprietary models in tackling complex visual question answering tasks.

- A noticeable directional bias was observed, as models showed a significant drop in performance (up to 15%) when answering questions about inverted flowcharts, suggesting a reliance on standard flowchart orientations.

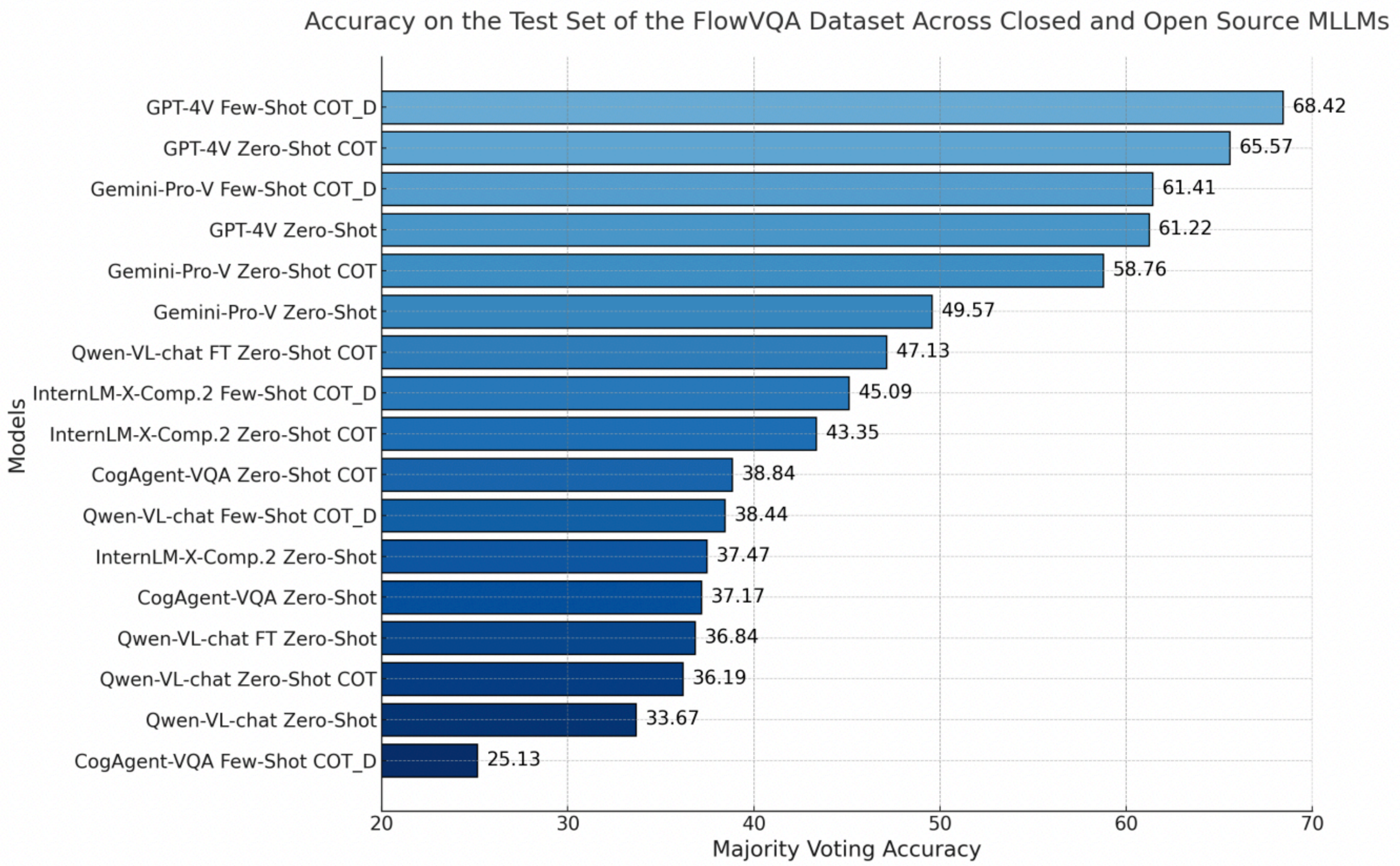


Figure 5: The horizontal bar chart shows the performance of FlowVQA dataset on various modelling strategies outlined in Section 3.

THANK YOU!

ACL 2024
Bangkok, Thailand